DOCUMENT RESUME

ED 173 446 TM 009 625

AUTHOR Paiva, Rosalia E. A.; Vu, Nu Viet
TITLE Standards for Acceptable Level of Performance in an Objectives-Based Medical Curriculum: A Case Study.
PUB DATE Apr 79
NOTE 16p.; Paper presented at the Annual Meeting of the American Educational Research Association (63rd, San Francisco, California, April 8-12, 1979)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Academic Standards; *Complexity Level; *Cutting Scores; Higher Education; *Mastery Learning; *Mastery Tests; *Medical Schools; Medical Students; Multiple Choice Tests; *Performance Based Education; Test Construction; Test Items; Test Reliability; Test Validity
IDENTIFIERS *Nedelsky Method

ABSTRACT

The Southern Illinois University School of Medicine applies mastery learning concepts to an objectives-based curriculum and uses the Nedelsky method to assess mastery of core objectives. Provisions are made for remediating deficiences. The Nedelsky technique is used for determining an acceptable level of performance in multiple choice examinations administered to medical students. Advantages of using the Nedelsky method are: (1) standards are based on content; (2) it helps faculty realize item and response option difficulties; (3) the technique is easily understood; and (4) data supporting the validity of the technique are already available. Shortcomings include: the likelihood of reduced rating reliability and validity when obtained from a single instructor; decreased emphasis on other important test characteristics; and the possibility of consistently underestimating an acceptable level of performance. It was concluded that the Nedelsky technique worked best in the following situations: several instructors who are familiar with the test content are available; the test contains a relatively large number of items; and the acceptable level of performance is adjusted to more than procedural considerations. (MH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED173446

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

Standards for Acceptable Level of Performance in an Objectives-Based Medical Curriculum: A Case Study

ROSALIA E. A. PAIVA, Ph.D., and NU VIET VU, Ph.D.

Southern Illinois University School of Medicine
Office of Curriculum Affairs and Educational Resources

Southern Illinois University School of Medicine graduated its first class in 1975. Since inception the School was committed to educational innovation. Accordingly, the systems approach to learning was adopted with objectives organized within an integrated framework and with progressive exposure to clinical learning activities.

Sound educational principles have been employed in all areas of curriculum development, instructional delivery, student progress monitoring and evaluation of student achievement. Among these principles, the following is of particular relevance to this paper: students are expected to master core objectives, and to meet predetermined acceptable levels of performance with provisions for remediating deficiencies when needed.

The teaching faculty, authors of the learning objectives, are responsible for setting acceptable levels of performance in advance of instruction. The difficulty of setting standards varies widely with the section of the curriculum, the type of competency evaluated and the clarity and specificity of the objectives.

This paper describes the use of the Nedelsky technique for determining the acceptable level of performance in multiple-choice examinations. Advantages and shortcomings of the technique are discussed.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Rosalia C.G. Paiva
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

Session Number: 18.14

Mailing Address:

Rosalia E. A. Paiva, Ph.D.
Associate Professor
Office of Curriculum Affairs
and Educational Resources
SIU School of Medicine
P.O. Box 3926
Springfield, IL 62708

TM009 625

Standards for Acceptable Level of Performance in an Objectives-Based Medical Curriculum: A Case Study

ROSALIA E. A. PAIVA, Ph.D., and NU VIET VU, Ph.D.

Southern Illinois University School of Medicine
Office of Curriculum Affairs and Educational Resources

The purpose of this paper is threefold: to call the attention of educators to efforts of an institution in applying mastery-learning concepts to an entire, objectives-based curriculum; to describe the use of the Nedelsky (1954) technique in setting standards for acceptable levels of student performance; and to gain insights from theoreticians and other practitioners in the panel and audience regarding the issues raised by this case study.

Southern Illinois University School of Medicine (SIU-SM) graduated its first class in 1975. Since inception the School was committed to educational innovation. Accordingly, the systems approach to learning was adopted with objectives organized within an integrated framework and with progressive exposure to clinical learning activities (Silber et al., 1978).

In embracing this objectives-based approach to instruction, the intent has been to maximize student development through the acquisition of a solidly organized core of medical knowledge, attitudes, skills and clinical problem-solving ability. Sound educational principles have been employed in all areas of curriculum development, instructional delivery, student progress monitoring and evaluation of student achievement. Among these principles, the following is of particular relevance to this paper: students are expected to master core objectives, and to meet predetermined acceptable levels of performance with provisions for remediating deficiencies when needed. Thus, the problem of allocating students to mastery states on the basis of criterion-referenced measurement constitutes a central issue in the evaluation system.

Since more than 8,000 learning objectives have been stated in the curriculum (Southern Illinois University School of Medicine, 1976) and evaluation of every student over every objective is not feasible, two alternatives are used. At the level of small units of instruction, such as learning modules and problem units, evaluation covers a sampling of objectives. For larger units of instruction, such as blocks or "sequences", evaluation pivots on more comprehensive or "cumulative" objectives which require mastery of prerequisite objectives and higher level cognitive processes. In either case, the aim is to determine individual mastery of competencies rather than class standing.

The teaching faculty, authors of the learning objectives, are responsible for setting acceptable levels of performance in advance of instruction by making judgments about what the student's behavior should be under specified conditions. In other words, faculty make judgments as to how well students must perform. The difficulty of setting standards varies widely with the section of the curriculum (basic sciences or clinical sciences), the type of competency evaluated (cognitive, attitudinal, clinical), and the clarity and specificity of the objectives. The less quantifiable the behavior called for by the objectives, the more subjective the judgment of the evaluators and hence, the more difficult the task of defining standards.

Paper presented in a symposium on "Setting Standards: Theory and Practice" at the annual meeting of the American Educational Research Association, San Francisco, April 8-11, 1979.

STANDARD SETTING: THE ACCEPTABLE LEVEL OF PERFORMANCE

More work has been performed in setting standards in those sections of the curriculum dealing with the cognitive domain. Written criterion-referenced or, more precisely, objectives-referenced tests and examinations are used in this area almost to the total exclusion of other evaluation methods. Exceptions comprise oral objectives-referenced examinations by one clinical department and oral quizzes for students who have been prescribed minor remediation.

Content validity is a primary concern in the process of developing examinations. Available relevant items are retrieved from item pools or new items are written, all keyed to the objectives to be evaluated. Once the selection of items meets test specifications, the instructor sets the cut-off passing score or, as referred to at SIU-SM, the acceptable level of performance (ALP) for the test.

At present the item formats most commonly used are the short-answer and completion format and the multiple-choice format, with the latter including both single best response and multiple response type items. Different methods of determining cut-off scores are employed for each item format. This paper deals specifically with the standard setting procedure applied to multiple-choice tests. This procedure is based on a technique developed at the University of Chicago by Nedelsky (1954). The method was selected over others because of its emphasis on item content, its apparent suitability to criterion-referenced testing, available validity data, and prior experience of some faculty with the method.

THE NEDELSKY TECHNIQUE

The Nedelsky technique originally was designed to determine course grades (mostly pass-fail) and make promotion decisions, by establishing absolute standards rather than standards relative to group performance. It applies only to multiple-choice examinations and focuses on the content of the items or, more specifically, on item difficulty. The technique has been used recurrently in medical education. It has been used at the University of Illinois College of Medicine with comprehensive examinations since 1962 (Note 1). Later the University of Missouri School of Medicine applied it to end-of-course pathology examinations (Taylor et al., 1971; Taylor and Reid, 1972). In 1975, SIU-SM adopted the procedure for written examinations in two of the four "sequences" comprising the curriculum.[1] Each of these three institutions made minor modifications in the technique apparently designed to fit the needs of their respective programs, in particular, their grading systems. The modifications have included: (a) different nomenclature for the cut-off score; (b) different difficulty codes assigned to the item response options by the University of Illinois; and (c) cancellation of a correction term in the original Nedelsky computation of the cut-off score by the three users. Because the basic procedure has remained essentially the same, the description of the technique as presently used at SIU-SM is presented first. Next, differences between SIU-SM and other users, including Nedelsky, will be highlighted.

The Nedelsky Technique as used at SIU-SM

The thrust of the method is to estimate the percentage of correct responses that would be answered by chance by the student meeting the minimum requirements of competency. This estimate is arrived at by analyzing each item and making judgments as to the most probable behavior of the student in regard to the various response options in the item. The task of the judge (coder or referee) is to identify which of the response options the "student with an acceptable knowledge for the material covered by the test item" should be able to rule out as definitely false. This judgment is then used to determine the acceptable passing index (API) for each item and the acceptable level of performance (ALP) for the test (Miller et al., 1978).

The acceptable passing index, or API, for an item is equal to the reciprocal of the number of options which the minimally competent student is not expected to identify as incorrect including the correct answer. In mathematical notation, the formula for computing the API for an item is

$$API = 1/(P-R)$$

where P is the number of response options in the item, and R is the number of response options that the minimally competent student should be able to rule out. Implicit in this definition of the API is the use of a binary difficulty code for the response options, namely, 0 (zero) for responses that are discarded as definitely false and 1 (one) for all remaining responses including the correct answer.

The API is taken as an estimate of the item difficulty and, accordingly, as the probability that a minimally competent student will choose the correct answer by chance. At variance with other standard-setting procedures which emphasize item content (Angoff, 1971; Ebel, 1972), this is a probability derived from the difficulty codes the judge assigns to the response options, rather than the judge's direct statement of the probability that the student will answer the question correctly. The lower the API, the more difficult the item, and vice versa. For items with five response options, the API may take any of the following values: .20, .25, .33, .50, or 1.0. For example, a five-choice item with no alternative being categorized as one that the minimally competent student should be able to discard as incorrect would have an API of 1/5 or .20 since all five options remain. Similarly, if the student is expected to rule out two options as incorrect, the API for the item would be 1/3 or .33, as there are three probable alternatives left from which to choose. Inspection of the possible values of the API for a five-option multiple-choice item makes it immediately apparent that they form a non-continuous scale with no values between .50 and 1.00.

The acceptable level of performance for the test, or ALP, is equal to the average of the API's over items. The ALP indicates the proportion of items the student must answer correctly in order to pass the test. Because of its definition, the ALP could be interpreted as an estimate of the total test guess score for the minimally competent student. At SIU-SM the ALP of a test usually is set by only one instructor. In the few occasions when more than one coder judges the items, the average of their ALP's becomes the ALP for the test. Nedelsky (1954) and the other two users of the technique--University of Illinois and University of Missouri--utilize multiple coders on a regular basis.

Differences Between SIU-SM and Other Users

Original Nedelsky Technique. In the original Nedelsky technique, the coders' mean ALP or "guess score" was increased by a term resulting from the product of a constant k and the standard deviation of the coders' API values. Nedelsky (1954) considered this standard deviation to represent "the most probable value of the standard deviation corresponding to the guess score" (p. 5) or, in other words, the standard deviation of the theoretical distribution of test scores of the "barely passing, borderline students". In SIU-SM terminology, Nedelsky's definition of the cut-off score would be expressed as

$$ALP = \overline{ALP} + k\sigma$$

where $\overline{ALP}$ represents the mean of all coders' ALP's. Nedelsky felt that because the mean of instructors' ratings of items or guess score represented an "absence of ignorance" standard, the $k\sigma$ correction was needed in order to decide in advance the percentage of "borderline" students that should fail the test. Thus, by setting k equal to either -1, 0, 1, or 2, he expected that on the average, 16%, 50%, 84% and 98% of these students would fail the test. Choice of intermediate values for k was also suggested.

Users of the Nedelsky technique in the field of medical education, including SIU-SM, have ignored this correction by tacitly assigning to k the value of zero. The University of Illinois seems to have utilized this or a similar adjustment factor of the ALP in the early sixties which later on was discontinued. In his excellent critical review of standard-setting procedures, Meskauskas (1976) took issue with Nedelsky's definition of the $k\sigma$ term and proposed an alternate adjustment to the determination of the cut-off point. He stated:

> "There appears to be no good reason to expect that the mean and S.D. of instructors' MPL [minimum passing level] values should be related to anything other than instructor/instruction variables. That being the case, use of a $K\sigma$ term seems unjustified although the basic concept of fine-tuning the setting of a minimum-passing point on the basis of a probability value may well have utility. A binomial or Bayesian approach would appear preferable to a normal model, however" (p. 136).

University of Illinois. There are two differences between the Nedelsky technique as utilized at SIU and the "modified" Nedelsky method used at the University of Illinois (U of I). They include:

1. U of I's use of difficulty codes 0, 1, and 2 for the various response options in the item, so that "...An option which a minimally passing student should recognize as incorrect is coded zero; one which a minimally passing student should not recognize as incorrect is coded two; all other options are coded one" (Smilansky & Guerin, 1976; p. 1). As the correct response is also given a code of two, the passing index of an item is equal to the sum of the difficulty codes, including the correct response, divided into two.

2. The second difference deals with nomenclature. U of I calls the cut-off score the "minimum pass level", or MPL, of the test (Note 1; Bobula & Standish, 1974).[2]

In regard to this nomenclature and difficulty codes, it is interesting to note that they were used by SIU-SM in the first experimental trial of the technique in early 1975. Two changes were implemented at SIU-SM as a result of this trial. First, the name of the cut-off score was changed to "acceptable level of performance", or ALP, to appease some teaching faculty who took issue with the expression "minimum pass level" as they tended to equate it with poor or mediocre student performance. Second, the use of U of I difficulty codes was discontinued because the faculty found them to be very confusing, particularly the use of the code 1 (one).

Validity of the Nedelsky Technique.

In regard to the validity of the results of the technique, several authors have provided data on: (a) the relationship of referees' a priori judgments of the response options with both empirical item difficulty and the choice of distractors by students at different achievement levels; (b) the relationship between acceptable passing indexes (or estimated item difficulty) and empirical item difficulty; and (c) comparison of students who scored above and below the cutting point on their performance on other achievement measures (Taylor et al., 1971; Taylor & Reid, 1972; Bobula & Standish, 1974; Smilansky & Guerin, 1976).

Reliability of the Nedelsky Technique

The effect of disagreement among "cutting score 'setters'" upon the reliability of standard-setting procedures has been duly raised by Brennan (1979, p. 39). Aside from the original work by Nedelsky (1954), three other studies performed with medical students have reported acceptable degree of agreement among judges (Note 1; Taylor et al., 1971; Taylor & Reid, 1972). However, highly divergent ratings of response options and wide range of coders' ALP's are sometimes obtained. Open discussion of departmental grading standards and objectives among judges has been recommended for these instances as a basis of reconciling differences (Note 1).

Failure to use the Nedelsky technique on account of extreme discrepancies among the estimates of six judges has been reported in only one study (Meskauskas & Webster, 1975). In this case, however, standards were being set not for the performance of students who had undergone instruction with the judges, but for a heterogeneous group of practicing physicians who were taking a recertification examination in internal medicine.

APPLICATION OF THE NEDELSKY TECHNIQUE AT SIU-SM

This section of the paper presents a summary of the application and use of Nedelsky's technique by various departments at SIU-SM.

Among the ten disciplines or departments which include multiple-choice items in their examinations, two have opted for the traditional method of

setting the ALP of a test by requiring students to obtain at least 70 or 80 percent of the test items correct. The remaining eight departments have adopted the Nedelsky technique. At present: (a) two disciplines use the technique without making any adjustment of the ALP; (b) four disciplines continue to use the technique but adjust the ALP after it is set; and (c) two disciplines have discontinued use of the technique. In the following paragraphs, the reasons which have led to adjustments or discontinuation of the use of the technique are presented.

One of the main reasons for adjusting the ALP prior to administering the test is faculty dissatisfaction with the calculated percentage of items students must answer correctly in order to pass the test. This occurs when faculty feel that the ALP is either too low or too high or, equivalently, too difficult or too easy in regard to level of training of the students. For example, if the preset ALP for a test of 100 items is found to be .40--meaning that the minimally competent student is expected to be able to answer correctly only 40 of the 100 items--, the faculty probably would judge the test to be too difficult. In addition, some faculty may become concerned about the negative psychological effect that such a low ALP may have on the students. The dissatisfaction leads to some kind of modification.

One type of adjustment consists of deleting some items and replacing them with items similar in content but of different degrees of difficulty (or API) in order to increase or decrease the total ALP for the test. Let us take the same example given above. If the total ALP of the test is .40 and the instructor decides to increase it (that is, make the test easier), some of the difficult items may be removed and replaced with easier items that have higher acceptable performance indexes (API).

In the second type of adjustment of the ALP before a test is administered, faculty dissatisfaction with their own ALP leads them to seek another coder's judgments. In many cases, this new coder is a physician resident who participates in the teaching of medical students. The midpoint between the instructor's and the resident's ALP is then used as the ALP for the test.

Both types of adjustments of the passing score originate from the instructor's perception that the test is either unreasonably difficult or unreasonably easy. Though subjective, the decision to adjust the ALP is based on the consideration of the relative importance of the content measured given the students' level of training, and the consequences of making incorrect decisions such as false retention or false advancement of students. Based on these considerations, making a test less difficult decreases errors of false retention, that is, remediating students over material that need not be learned 100 percent. On the other hand, by making the test more difficult the instructor minimizes the risk of passing students who have not yet achieved mastery of the objectives.

Aside from instructor dissatisfaction with the calculated ALP and the lengthy process of determining the API for the test items, the main reasons which have led some faculty to the rejection of the technique are the restricted range of probabilities for each item and the effect of this constraint on the ALP, especially in short tests. For example, one discipline only evaluates students on subtest performance rather than on total test performance. In other words, in order to pass, students must score at or above the cut-off score on each subtest even when their score on the entire test is higher than the total ALP. If a subtest has only three five-choice items where each item

has an API of .33, the calculated ALP for the subtest would be .33 + .33 + .33, that is, one question out of three. In this case, as Millman (1973) also has observed, the students who answer correctly one item would pass even though they may choose distractors for the other items which the minimally competent student should be able to rule out.

This problem is particularly important in medicine where the wrong choice of a treatment modality, for example, may be the cause of an array of negative consequences. As Millman (1973) states in his criticism of probabilistic models for setting proficiency standards, including the Nedelsky technique:

> "In fields like medicine, the system permits a student who knows the answer to a few items to select options describing dangerous practices in perhaps many more items and yet not fail the test" (p. 208).

The problem can be circumvented in several ways. One of them, suggestd by Millman (1973), is to consider the difficult-to-discriminate options as correct options; this implies penalizing students only when the incorrect-harmful option is chosen. Another alternative is to assign different weights to options considered (a) correct, (b) incorrect but not harmful, and (c) incorrect but harmful. A third possibility is to give students partial credit for choosing options which are correct but incomplete, or options which are nearly correct. Although these suggestions may solve the problem of passing a student for not choosing dangerous practices, they do not take into consideration the difficulty of the response options. Better yet would be to develop a standard-setting procedure which would consider both the difficulty and degree of "correctness" of the response options. Such procedure would employ criteria different than those proposed in Ebel's model (1972), with instructor's judgments targeted to response options rather than to entire items.

All of the adjustments of the ALP described above are made prior to the administration of the test. Once the test is administered, further adjustments are possible based on the results obtained from an item analysis. These a posteriori adjustments are used both for tests where the cut-off score had been set following the traditional method and for tests where the ALP had been determined by the Nedelsky technique. The adjustment usually consists of deleting items that instructors interpret as poor or ambiguous based on the item statistics. The ALP is then reset based on the shortened test.

In summary, most of the disciplines which have adopted the Nedelsky technique adjust the obtained acceptable level of performance of a test mainly because of the faculty's uneasiness with and poor confidence in the set standard. Most of the adjustments are made prior to the test administration by taking into consideration factors other than the difficulty of the items alone. These factors include the relative importance of the objectives measured, the students' level of training, the consequences of making false advancement and false retention decisions, and psychological consequences on the students. Shortcomings encountered with the technique when applied to subtests consisting of few items account almost totally for discontinuation of use by at least one department.

DISCUSSION

Overall, SIU-SM is satisfied with the Nedelsky technique as a procedure to set the acceptable level of student performance on multiple-choice examinations. Data presently at hand suggest that advantages may outweigh shortcomings.

One of the major advantages of the technique is that standards are established based on consideration of content. In an objectives-based system designed around the concept of mastery of specified competencies, content seems the most relevant and appropriate criterion on which to found decisions. This argument is further convincing when one heeds errors introduced in test development. Could practitioners place more confidence in their exams, they might be inclined to more readily accept standards based on criteria somewhat foreign to them such as those established within a statistical framework.

Another advantage that the Nedelsky method shares with other content-bound procedures of setting standards (Angoff, 1971; Ebel, 1972) is that it helps conscientious faculty to realize the level of difficulty of items and act accordingly.

Furthermore, at variance with the Angoff (1971) and the Ebel (1972) techniques which determine the test ALP mainly from the overall difficulty of the item, the Nedelsky technique makes the instructor aware of the difficulty of the response options. In so doing, it calls the attention of the instructor to the student's dilemma when confronted with different types of distractors. In regard to these techniques, Meskauskas (1976) states:

> "...It appears likely that, in contrast to the Nedelsky method, the Ebel method"... [and possibly the Angoff method as well] ..."would allow the rater to ignore some of the fine discriminations that an examinee needs to make and would result in a standard that is more difficult to reach" (p. 138).

When faced with a multiple-choice item, the student's task entails selecting the correct response from among several alternatives. In other words, the student is expected to respond to one stimulus among several others. Given present widespread commitment to the use of multiple-choice examinations, instructor awareness of all the stimuli presented to the student as required by Nedelsky's method might provide a more informed basis for setting standards for criterion-referenced measurement.

Still another advantage of the Nedelsky technique is that it is easily understood by instructors and therefore easy, albeit somewhat tedious, to apply.

Lastly, data supporting the validity of inferences based on this standard setting procedure are available from independent studies which have already been reported (Taylor et al., 1971; Taylor & Reid, 1972; Bobula & Standish, 1974; Smilansky & Guerin, 1976).

Considerations regarding the validity as well as the reliability of the technique, however, lead us to a characteristic of the method which becomes a shortcoming when applied within a program such as SIU-SM. The characteristic referred to here is that the technique calls for multiple judges.

As is the case with most, if not all, methods of establishing cut-off scores, the validity and dependability of the standard obtained with Nedelsky's model seems far too dependent on the judgment of the instructor, including his/her experience, limitations and wisdom (Shepard, 1976). Thus, the acceptable passing index (API) for an item and therefore the acceptable level of performance (ALP) for a test may be inordinately affected by the coders' ability, or lack of ability, to put themselves in the place of the student when confronted with the item. Since by definition the instructor is a content expert and possesses more knowledge of the subject than the minimally competent student, the difficulty of the items may be underestimated and a hardly-achievable passing score be set (Levine & Forman, 1973). The opposite case is also likely to occur; that is, instructors may resort to a benefit-of-the-doubt type of judgment behavior which can result in artificially inflating the estimated difficulty of the test with the corresponding decrease in the value of the ALP. In both these cases, decisions based on the standards would lack validity as misclassification of students would take place. Occurrences of the first instance leading to unduly high cut-off scores because of judges' inexperience have been observed at two institutions.[3]

The available data on the validity of the technique when several coders are used, and the finding of discrepancies between expected and observed performance when a single coder or inexperienced judges are used, all indicate that the technique would best be applied as originally designed by Nedelsky. That is, the use of multiple experienced coders seems to be a sine qua non condition to obtaining reliable and valid cut-off scores.

In his review of the technique, Meskauskas (1976) concludes:

> "The impact of item-by-item judgments on failure rate is very clear".

Therefore,

> ..."The procedure used to arrive at the [ALP] must ensure a broad base of judgment and consensus so that a common standard is, in fact, achieved" (p. 137).

The availability of only one judge for coding some of the examinations is an empirical problem encountered at a relatively small school like SIU-SM. This is a greater problem with sections of the curriculum involving highly specialized content areas.

Solutions sought by the faculty to their dissatisfaction were mentioned earlier. In order to increase validity of their judgment and decrease error, some instructors adjust their predetermined standards by modifying or deleting items, others by asking for the judgment of residents in their services, and still others by resetting the cut-off score, after administering the test, on the basis of how students have performed.

An alternative to these solutions is the use of item performance data from previous groups of instructed and non-instructed students when available. Assuming similar initial skills, motivation and knowledge in successive groups of students and no variability in the quality of instruction, such data can serve as a guide for the instructor's judgments as to standards setting. In addition, as pointed out by Hambleton and his colleagues (1978), this information might also be used to determine construct validity of the scores derived from the test. Shepard (1976) has recommended not only that several expert judges be involved in the process of setting standards, but also that these expert judges be provided with normative data in their deliberations. To avoid misinterpretations, she adds:

> "Of course, the norms are not automatically the standards, Experts still have to decide what 'ought' to be; but they can establish more reasonable expectations if they know what current performance is than if they deliberate in a vacuum" (p. 11).

Aside from the lack of multiple judges, a second shortcoming of the Nedelsky technique when applied at SIU-SM is its lack of consideration of factors other than the difficulty of item response options. One important factor not considered by the procedure is the consequence of decisions.

The problem of setting standards for acceptable levels of performance in an objectives-based system--or, in other words, the problem of assigning individuals to mastery states in regard to prescribed competencies--cannot be disassociated from the consequences of the decisions made once the student behavior has been "measured". At SIU-SM financial consequences and, to some extent, psychological consequences of the decision to retain or advance students are secondary to educational considerations. Thus, more rigorous standards are set for the assessment of objectives which either are prerequisite to other learning, deal with core competencies, or are defined to represent "cumulative" or more comprehensive levels of achievement.

Superceding all of these considerations, however, is the concern for later costs to society of a non-proficient individual graduating from medical school. Thus, because we are a professional school educating people to become physicians, there is more concern about passing standards leading to false positive than to false negative errors. In other words, generally the faculty would rather set standards and make decisions which might result in more proficient students having to remediate for objectives already achieved than erroneously passing fewer apparently proficient students who in fact should not pass. It is mainly this belief and ultimately the regard for the role that the physician plays in society, that justifies the resetting and adjustments of the acceptable level of performance on the part of instructors.

A third shortcoming of the technique relates to the probability levels of obtaining an item correct being limited to discrete values between .20 and .50 and no values between .50 and 1.00. As a result of these limits on the probability scale, there may be a consistent underestimate of the number of questions a student must answer correctly in order to pass the test. The possibility of this underestimation is more accentuated when the Nedelsky technique is applied in short tests. In this case, as the variability of the difficulty among items is more affected than in long tests by the restricted range of probabilities, the ALP may be further underestimated.

This issue should be interesting to pursue, particularly in comparison to other content-bound standard-setting procedures. As was pointed out earlier, the focusing of other techniques such as Ebel's on the probability for the entire item, has been questioned as a source of overestimate of the passing standard in this method.

In regard to the technique limiting the process of setting standards to the difficulty of the response options, factors not considered by the procedure include the relative importance of the behavior measured, classification errors that may result in false advancement or false retention of students, the length of the test, and the estimate of prior performance. Although several models which are designed to set acceptable standard of performance have considered the above factors in various combinations (Davis & Diamond, 1974; Ebel, 1972; Emrick, 1971; Kriewall, 1972; Millman, 1972 and 1973; and Novick & Lewis, 1974), presently they could not be applied within our curriculum for several reasons. These are: (a) some of the assumptions on which the models are based could not be met within our system, (b) there is not sufficient validity data obtained on the models, and (c) the procedures called for by the models do not seem practical and easy enough to use in order to insure adequate application and accurate decisions.

CONCLUSIONS

In summary, use of the Nedelsky technique at SIU-SM has revealed three possible shortcomings, the significance of which needs to be evaluated. These are: (a) the likelihood of reduced reliability and validity of the ratings when obtained from a single judge; (b) the emphasis of the technique on the difficulty of the response options to the exclusion of other considerations (i.e., importance of content, misclassification consequences, length of test); and (c) the likelihood of a consistent underestimate of the acceptable level of performance.

In spite of these shortcomings, the Nedelsky technique seems to be the most appropriate technique to apply within our objectives-based curriculum. The technique not only uses content as a main criterion for setting the acceptable level of performance, but also is practical and easy to apply, has available validity data, and makes faculty aware of the level of difficulty of their tests, including item response options.

Acknowledging shortcomings, the technique seems to be best applied to testing situations where: (a) multiple judges familiar with the test content are available; (b) the test includes a reasonably large number of items; and (c) the obtained acceptable level of performance is adjusted on the basis of considerations other than those dictated by the procedure.

FOOTNOTES

[1]Evaluation procedures in the other two sequences do not lend themselves to the use of Nedelsky's technique.

[2]The name of the cut-off score at the University of Illinois has recently been changed to "minimal acceptable performance", or MAP (Smilansky & Guerin, 1976).

[3]A few years ago SIU-SM faculty were asked to use the Nedelsky technique and establish an ALP for National Board of Medical Examiners (NBME) Part II Subject Tests. These consist of subtests in each of the major clinical disciplines. Two judges (teaching faculty) of each department met with a member of the educational staff who assisted with the coding of the items and who periodically reminded the judges that the coding of response options was to be based on the expected performance of the SIU-SM student meeting minimal competence level.

A common observation on the part of the staff was that the judges' reaction to most items consisted of identifying the correct answer and then making a somewhat global judgment of the item relative to the ease or difficulty they themselves experienced in answering the item. Furthermore, it appeared that they could not detach themselves from their own level of competence and knowledge. Calculation of the API for each item and ALP for the test confirmed this impression in that the result was an underestimate of the difficulty of the test and, therefore, a spuriously high passing level even in comparison with that set by NBME.

Had the standards generated by the SIU-SM faculty been applied to the results of the examination, most students in the United States would have failed the test. In this example, the discrepancy between judges' expectations and student performance was due mainly to the lack of experience with the method.

It is interesting to note that a similar observation had been made at another institution. One of the first attempts of the University of Illinois to apply the Nedelsky technique to comprehensive exams in 1963 also produced recommended standards that resulted in a failure rate "excessive by any criterion" (Note 1). When the University of Illinois compared these results with those of a different group of judges who had used the method in three previous occasions, it was concluded that the additional experience had allowed this latter group to develop standards in which they could place better confidence.

REFERENCES

Angoff, W.H. Scales, norms, and equivalent scores. In R.L. Thorndike (Ed.), Educational Measurement. Washington, D.C.: American Council on Education, 1971.

Bobula, J.A., & Standish, M. Minimum Pass Level Study. Report to the Faculty, Center for Educational Development, University of Illinois College of Medicine, 25-31, 1974.

Brennan, R.L. Some Applications of Generalizability Theory to the Dependability of Domain-Referenced Tests. ACT Technical Bulletin No. 32. Iowa City, Iowa: American College Testing Program, April 1979.

Davis, F.B., & Diamond, J.J. The Preparation of Criterion-Referenced Tests. In C.W. Harris, M.C. Alkin & W.J. Popham (Eds.), Problems in Criterion-Referenced Measurement. Los Angeles: UCLA Graduate School of Education, Center for the Study of Evaluation, 1974.

Ebel, R.L. Essentials of Educational Measurement. Englewood Cliffs, NJ: Prentice-Hall, 1972.

Emrick, J.A. An Evaluation Model for Mastery Testing. Journal of Educational Measurement, 8:321-326, 1971.

Hambleton, R.K., Swaminathan, H., Algina, J., & Coulson, D.B. Criterion-Referenced Testing and Measurement: A Review of Technical Issues and Developments. Review of Educational Research, 48:1-47, 1978.

Kriewall, T.E. Aspects and Applications of Criterion-Referenced Tests. Downers Grove, IL: Institute for Educational Research, April 1972. (ERIC Document Reproduction Service No. ED 063 333)

Levine, H.G., & Forman, P.M. A Study of Retention of Knowledge of Neurosciences Information. Journal of Medical Education, 48:867-869, 1973.

Meskauskas, J.A. Evaluation Models for Criterion-Referenced Testing: Views Regarding Mastery and Standard Setting. Review of Educational Research, 46:133-158, 1976.

Meskauskas, J.A., & Webster, G.D. The American Board of Internal Medicine Recertification Examination: Process and Results. Annals of Internal Medicine, 82:577-581, 1975.

Miller, H.G., Williams, R.G., & Haladyna, T.M. Beyond Facts - Objective Ways to Measure Thinking. Englewood Cliffs, NJ: Educational Technology Publications, Inc., 1978.

Millman, J. Tables for Determining Number of Items Needed on Domain-Referenced Tests and Number of Students to be Tested (Technical Paper No. 5). Los Angeles: Instructional Objective Exchange, April 1972.

Millman, J. Passing Scores and Test Lengths for Domain-Referenced Measures. Review of Educational Research, 43:205-216, 1973.

Nedelsky, L. Absolute Grading Standards for Objective Tests. Educational and Psychological Measurement, 14:3-19, 1954.

Novick, M.R., & Lewis, C. Prescribing Test Length for Criterion-Referenced Measurement. In C.W. Harris, M.C. Alkin, & W.J. Popham (Eds.), Problems in Criterion-Referenced Measurement. Los Angeles: UCLA Graduate School of Education, Center for the Study of Evaluation, 1974.

Shepard, L.A. Setting Standards and Living With Them. Paper presented at the 1976 meeting of the National Council on Measurement in Education, San Francisco, April 1976.

Silber, D.L., Williams, R.G., Paiva, R.E.A., Taylor, D.D., & Robinson, R. The SIU Medical Curriculum: Systemwide Objectives-Based Instruction. Journal of Medical Education, 53:473-479, 1978.

Smilansky, J., & Guerin, R.O. Minimal Acceptable Performance Levels for Criterion-Referenced Multiple Choice Examinations and Their Validation. Paper presented at the 1976 meeting of the American Educational Research Association, San Francisco, April 1976. (ERIC Document Reproduction Service No. ED 123 255)

Southern Illinois University School of Medicine. Curricular Objectives - 1976. Carbondale, IL: Author, 1976.

Taylor, D.D., & Reid, J.C. Criterion-Referenced Evaluation of Student Performance. Journal of Medical Education, 47:970-971, 1972.

Taylor, D.D., Reid, J.C., Senhauser, D.A., & Shively, J.A. Use of Minimum Pass Levels on Pathology Examinations. Journal of Medical Education, 46: 876-881, 1971.

REFERENCE NOTES

1. Setting Standards of Competence. The Minimum Pass Level. The Evaluation Unit, Center for the Study of Medical Education, University of Illinois College of Medicine, January 1967.